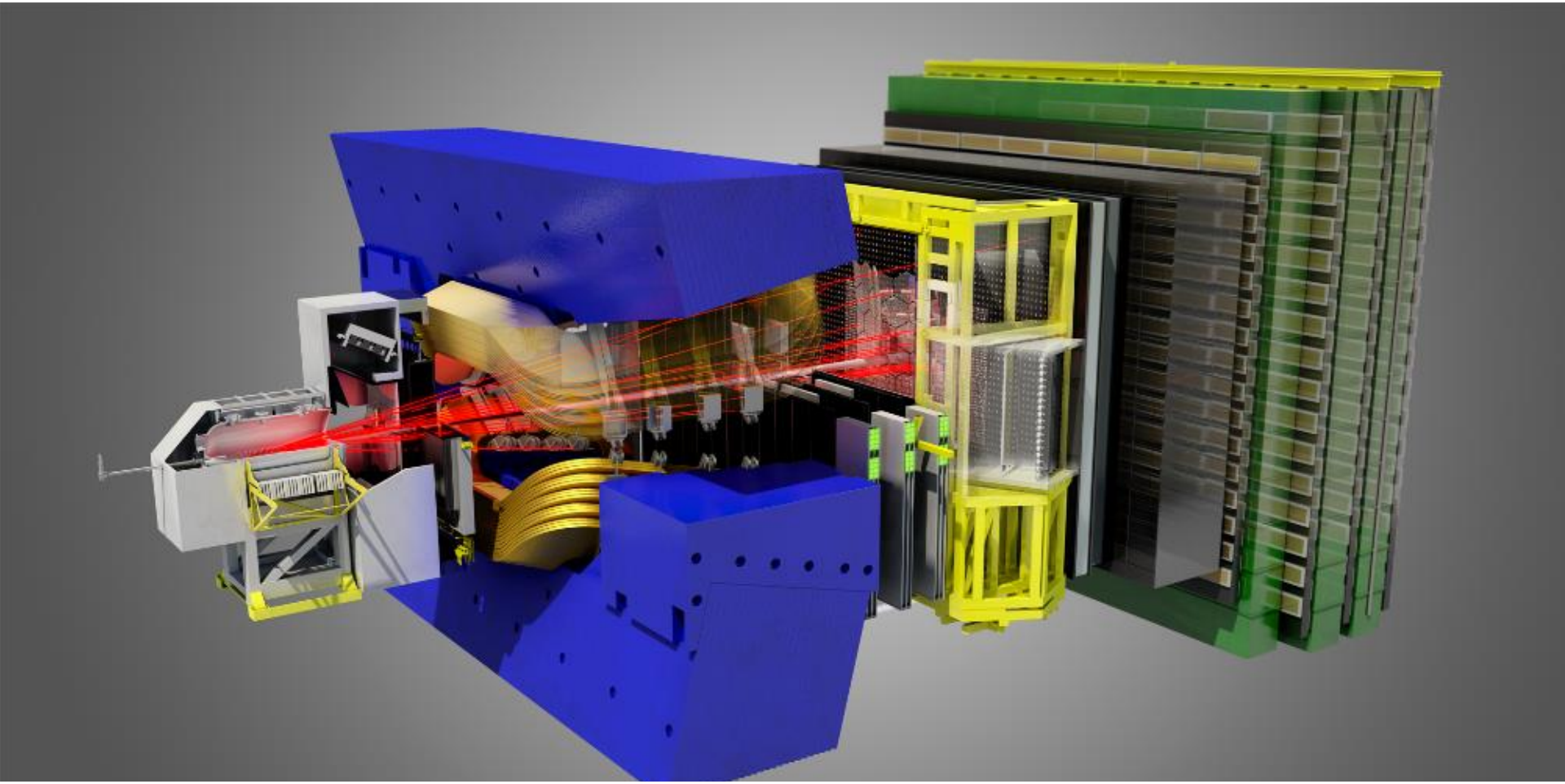


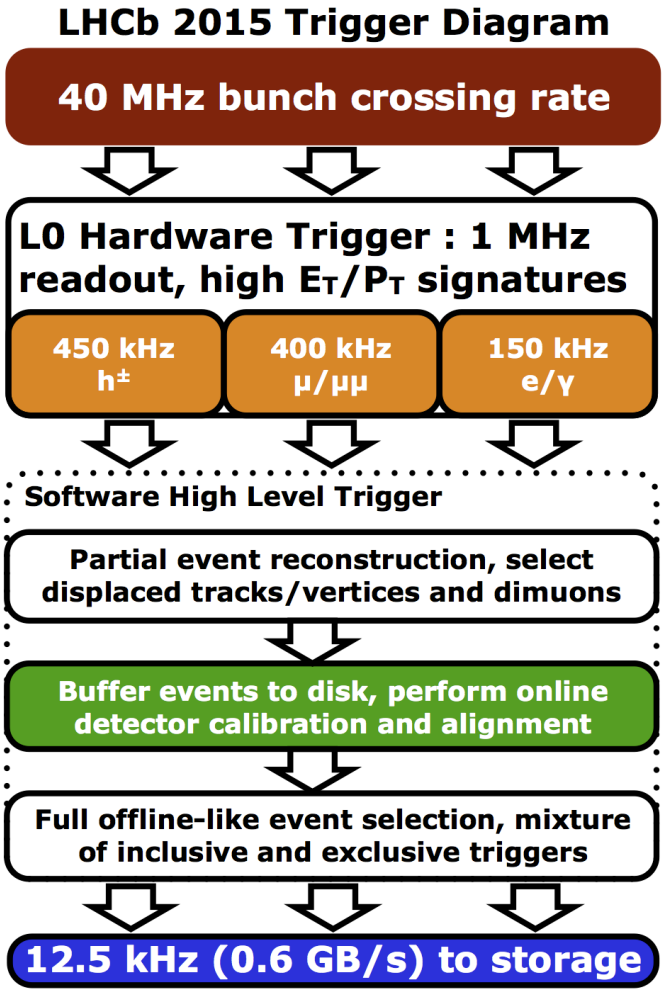
LHCb: Online Compute system for Run 3

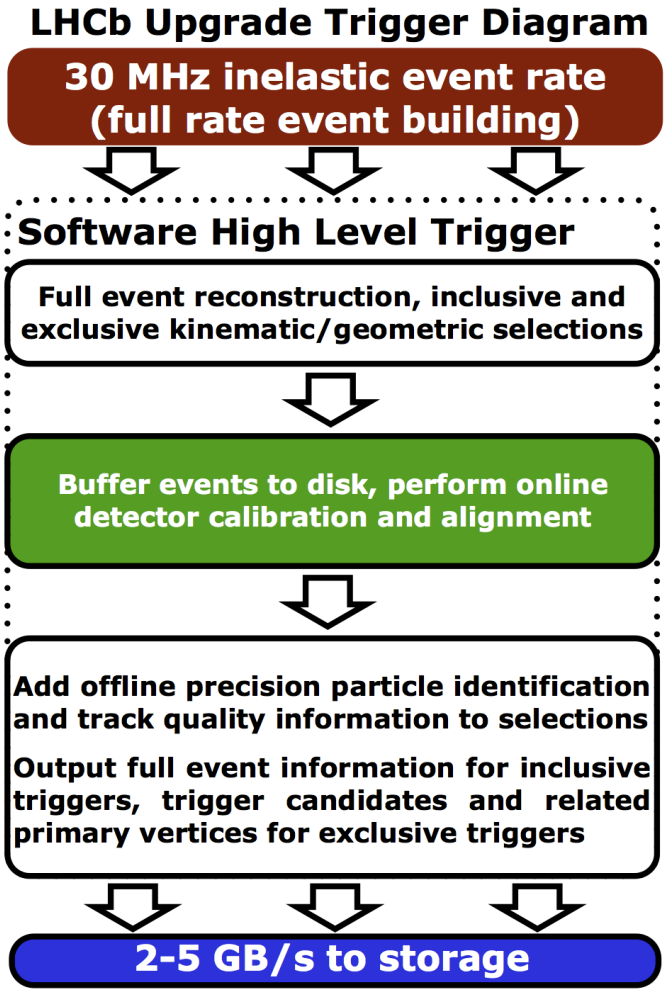
Niko Neufeld
niko.neufeld@cern.ch

Outline

- Readout system
  - Slow & fast control
  - Optical links
  - Readout board
  - Event building
  - Data centre

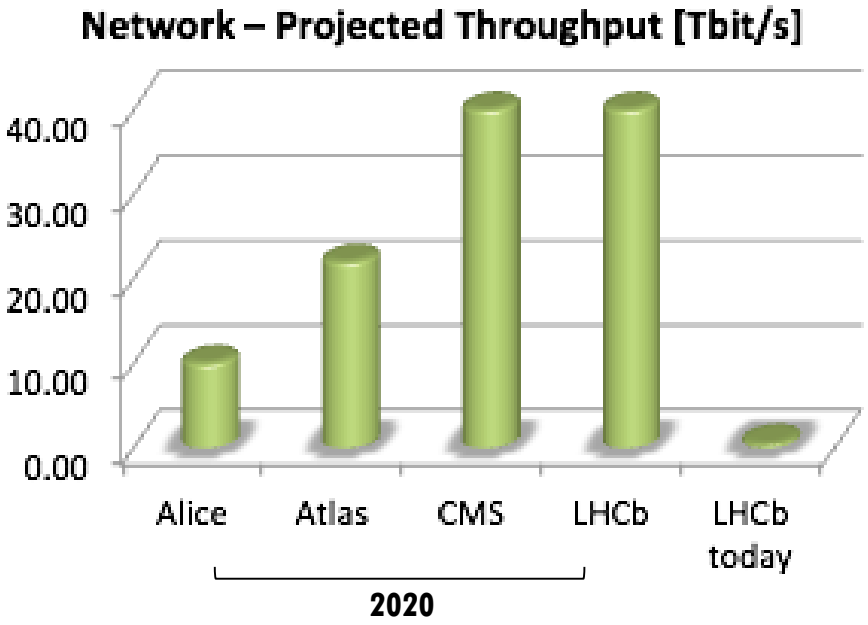
LHCb Trigger in from Run2 to Run3





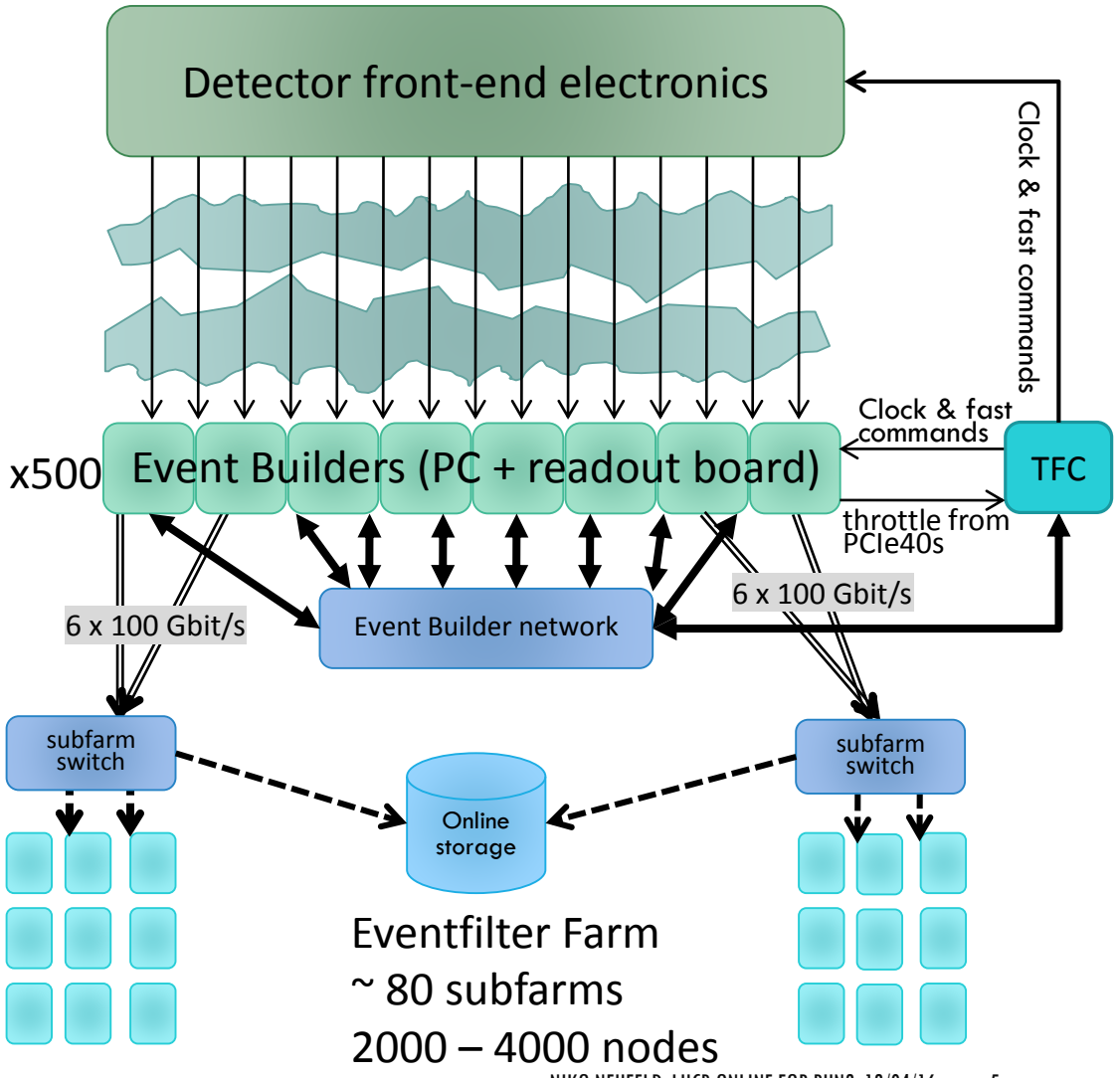
Run3 upgrade

- Filter farm will need to handle:
  - **Event size** (~130 kB) (@ 30 MHz)
  - Larger **event rate** (40MHz == LHC bunch crossings per second)
- New challenges for DAQ & High-Level Trigger



Run3 Online System

- Dimensioning the system:
  - **~10 000** versatile links
    - detector to surface **(~350 m)**
    - up to **4.8 Gbps/link**
  - **~500** readout nodes (up to ~24 links/node)
  - **~40 MHz** event rate
  - **~130 kB** event size (@ 27 MHz)
- High bisection bandwidth in event builder network
  - ~32 Tb/s aggregate bandwidth
  - Leverage 100G Lan technologies
- Global configuration and control via ECS subsystem
- Global synchronization via TFC subsystem

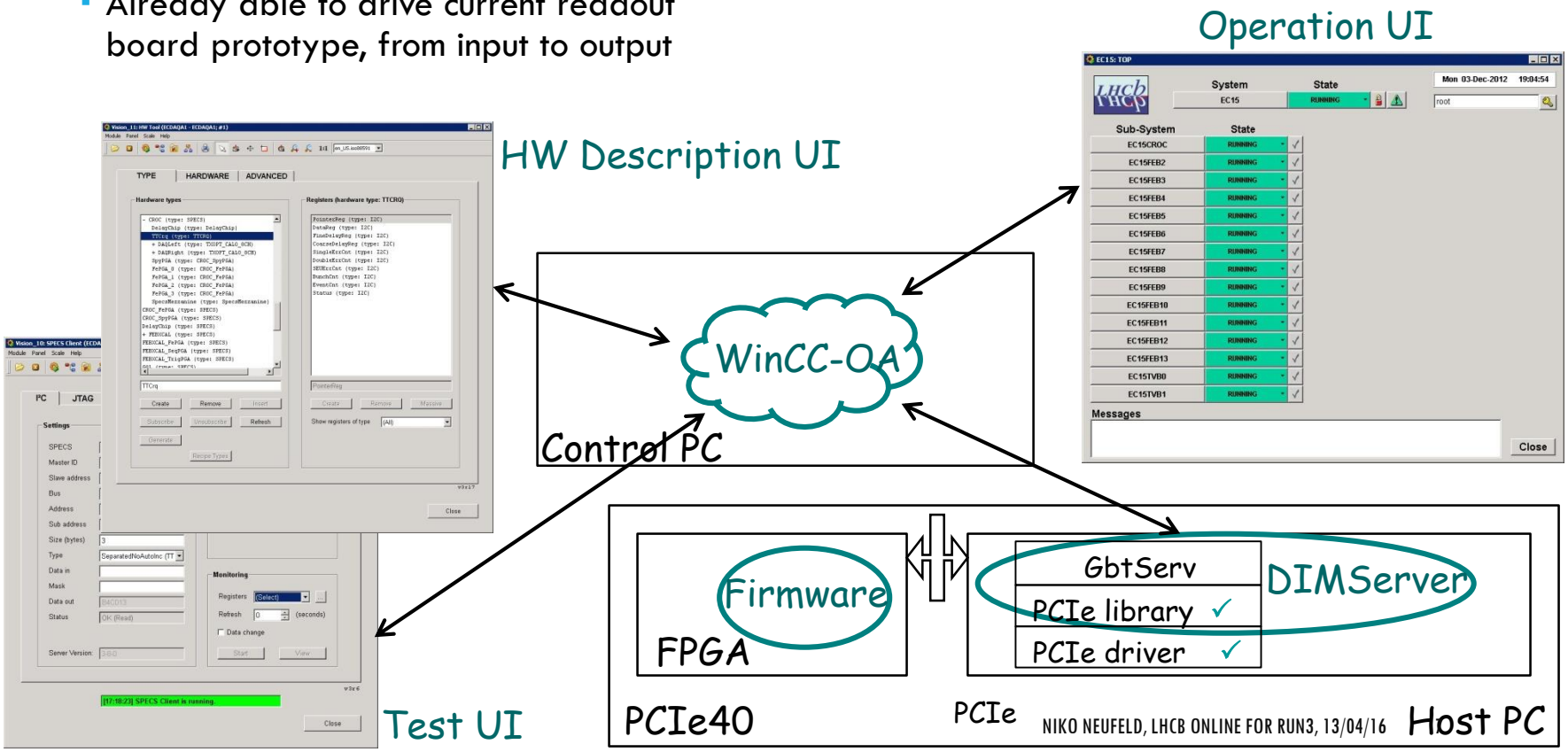


Numbers

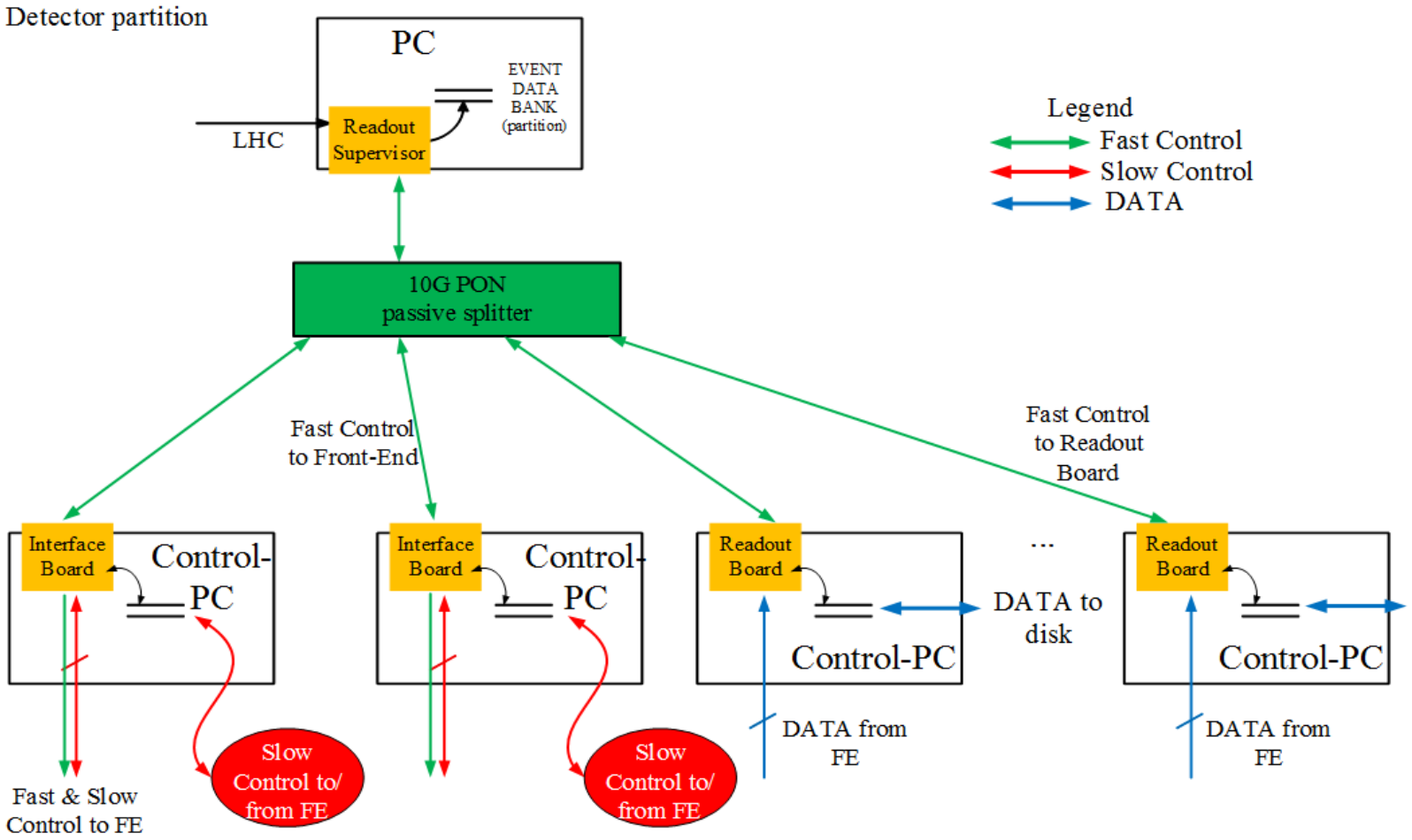
- Event-size (@ 2x10^33) ~ 130 kB
- Readout-rate 40 MHz
- 500 event-builder nodes
- Between 1000 and 4000 event-filter nodes
  - Dual-socket, accelerator to be decided
- 500 port minimum event-building network
  - TDB: OPA, IB, Ethernet
- 1500 – 4500 port filter network
  - Ethernet?
- New data-centre
  - 4000 rack-units
  - 2 MW max
- 50 to 100 real nodes for “slow” and “fast” control
  - Using PCIe40 cards
- Rest of control-system on virtual machines as today
- Local storage on each filter-unit at least 20 TB → will depend on disk-technology
- Central buffer storage ~ 1 to 2 PB
- ~ 10000 uni-directional fibres for DAQ (4.8 Gbit/s)
- ~2000 fibre-pairs for ECS/TFC (GBT)

Experiment Control System for Run 1, 2, 3, 4, 5, 6 …

- Controls and monitors <u>all subsystems</u>
  - DAQ, TFC, HLT, farm…
- Continuity from current implementation
  - JCOP / DIM / WinCCOA / SMI++ / Recipes
- Already able to drive current readout board prototype, from input to output
- Frontends rely on GBT-SCA hardware by EP-ESE
- Low-level components are being implemented



TFC architecture



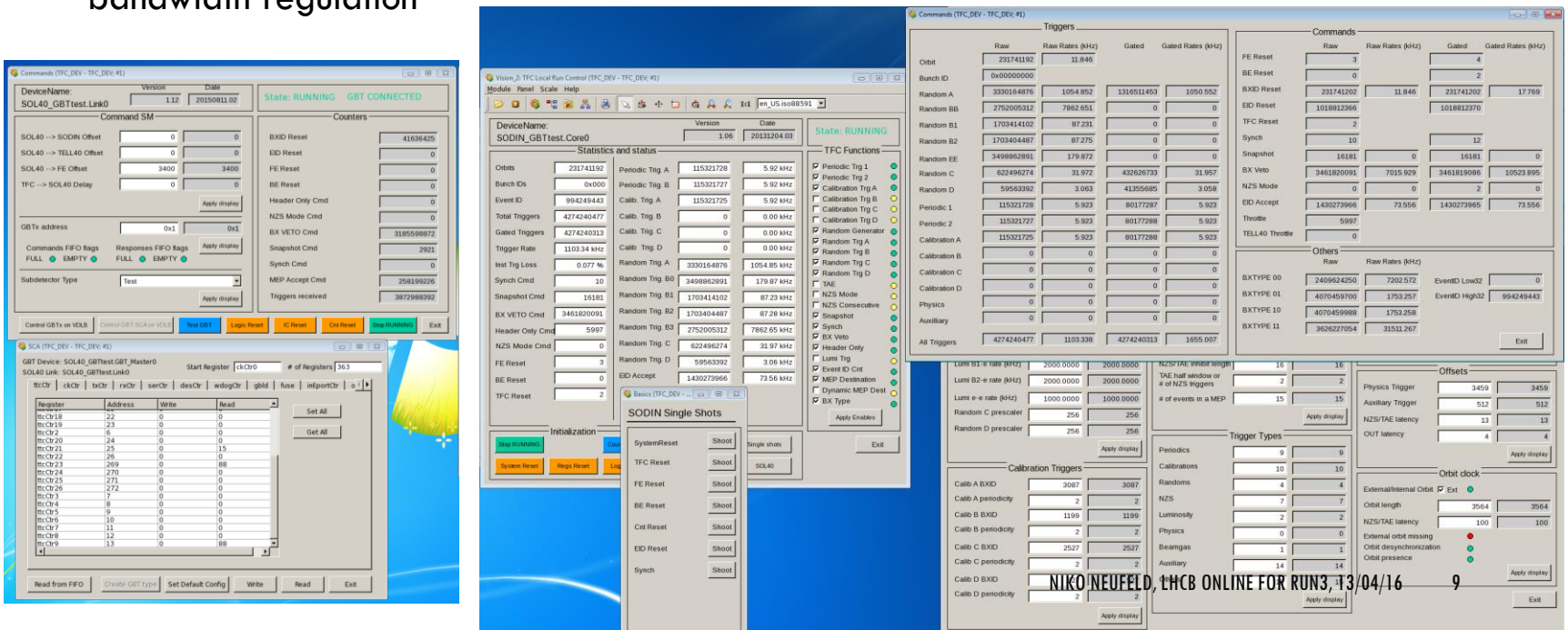
TFC (Timing & Fast Control)

Current status

- Already integrated in firmware
- Uses same readout board hardware as the DAQ (PCIe40)
- Can send fast commands to frontends
  - SciFi, UT, Muon ASICS already being tested
- Programmable internal throttle for bandwidth regulation

Ongoing work

- PON (Passive Optical Network) technology integration (with EP-ESE)
- Clock phase tests on readout board (with CPPM)
- Continue feedback and compliance testing with frontend experts



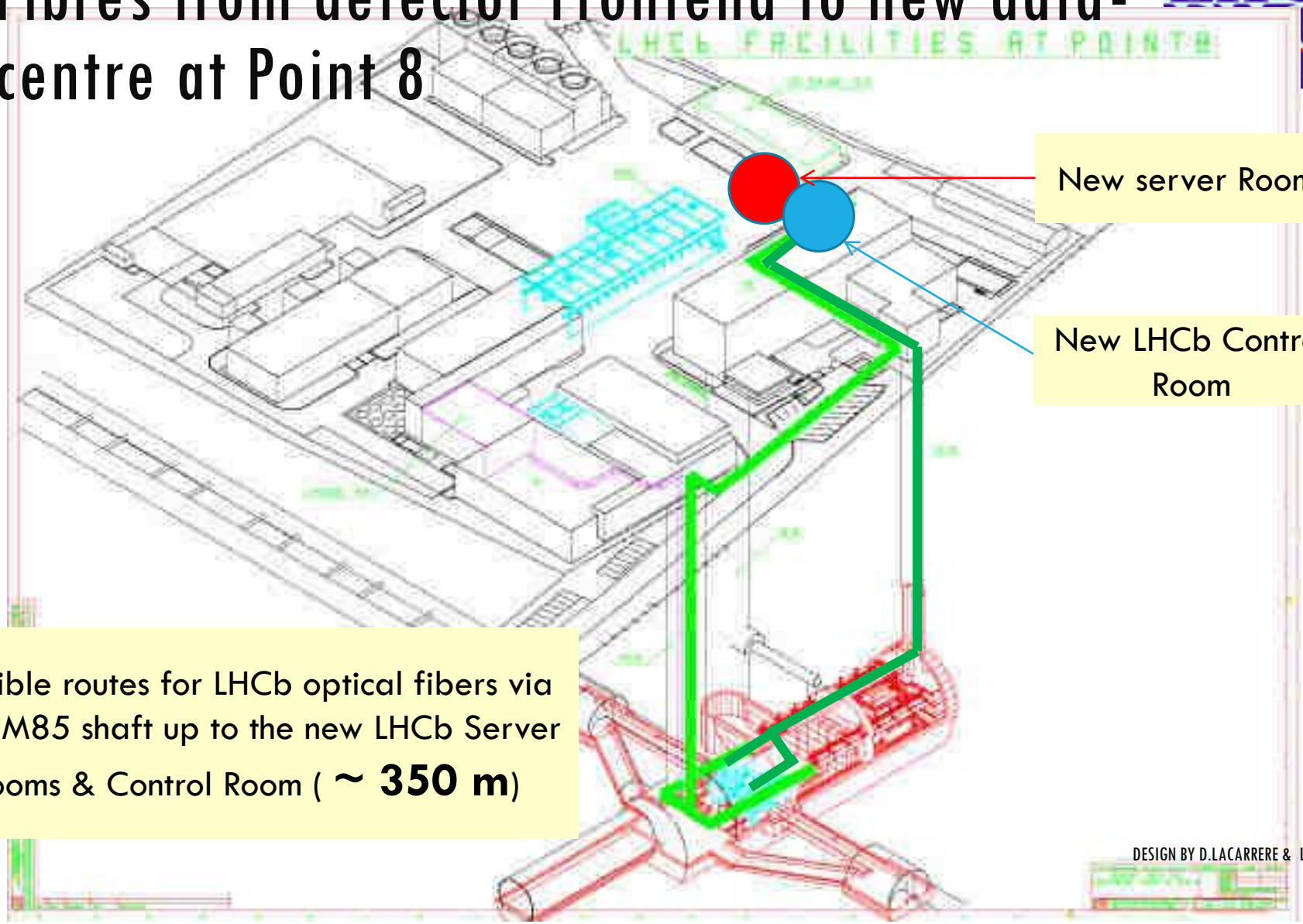
DAQ cost optimisation

Main cost driver (apart from CPU) is the number and length of fast interconnects (like in HPC!)

Versatile links are there and optical anyhow

Most compact system has everything: event-builder, controls, TFC and farm in one place → new data-centre at the surface and run versatile links to the surface

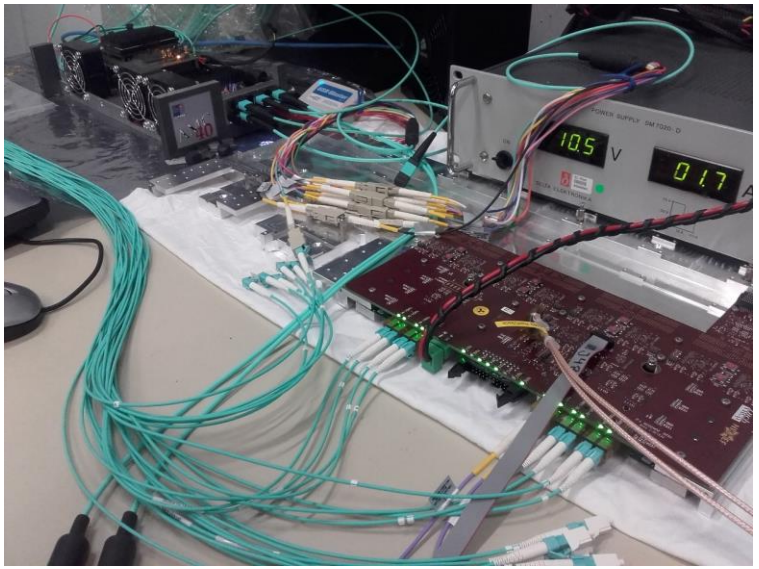
Fibres from detector Frontend to new data-centre at Point 8

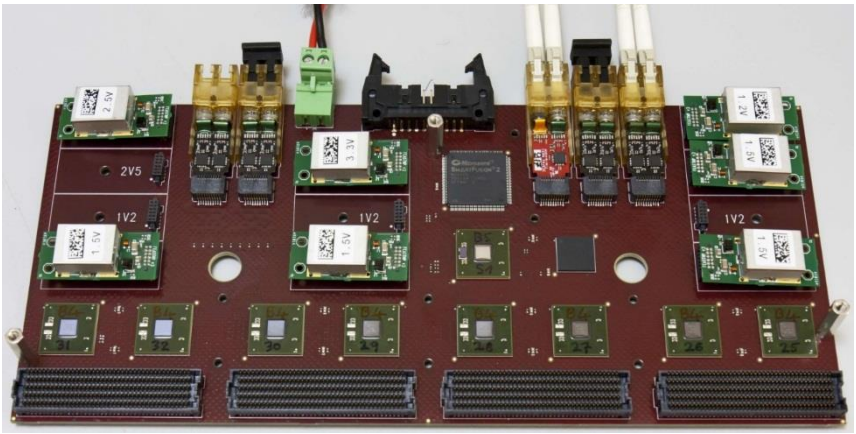


Possible routes for LHCb optical fibers via the PM85 shaft up to the new LHCb Server Rooms & Control Room (**~ 350 m**)

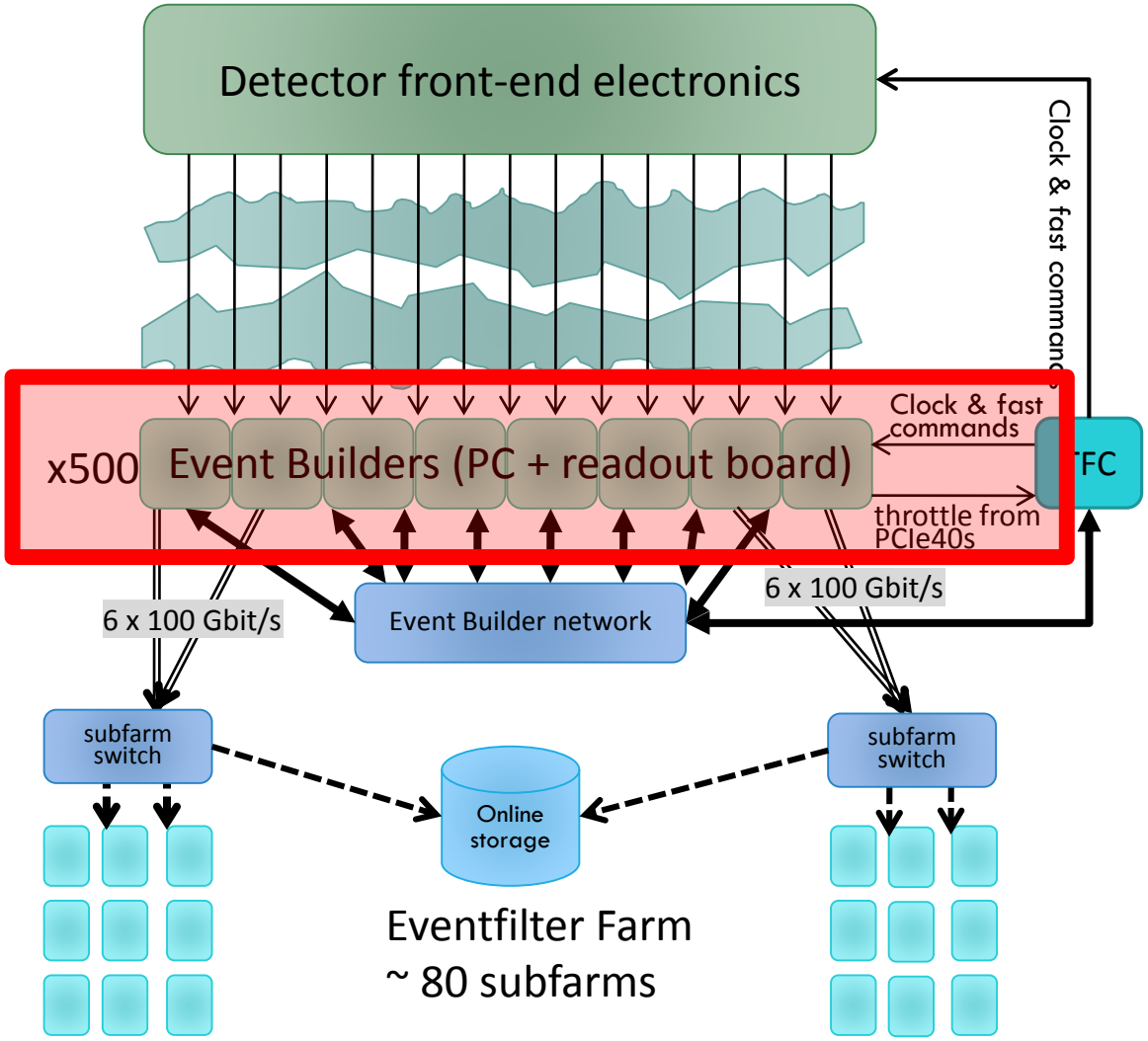
Long-distance optics

- Counting room on surface
  - Power, cooling, space constraints in underground area
  - ~350 meter distance over OM3 MMF
- Based on EP-ESE technology
  - Rad-hard Versatile Link on frontends
  - Initially qualified for ~100m
- Fiber infrastructure by EN-EL
  - Pilot installation at end 2014
- Loopback tests in 2015
  - ~9 months, ~700 meters
  - Avago MiniPOD transceivers
  - Bit Error Rate < 10^{-18}
  - Full system equivalent: < 5 errors/day
  - ✓ LHCC milestone
- Continued tests in 2016
  - Versatile TX on frontend prototype
  - MiniPOD RX on readout board prototype



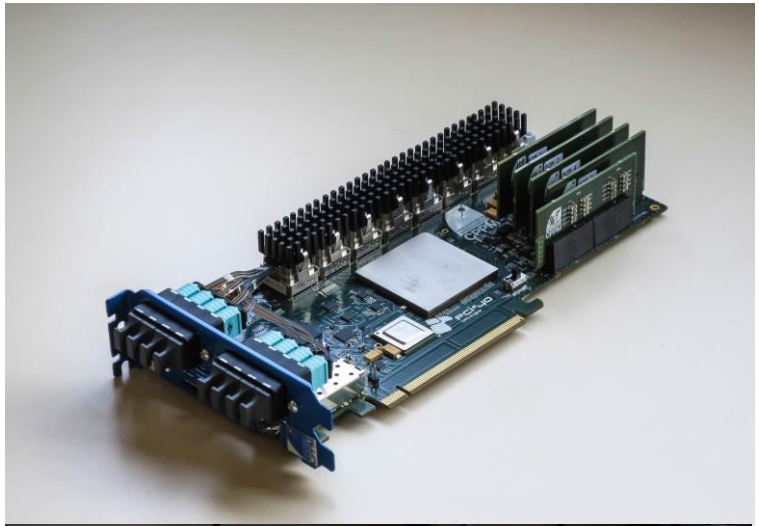


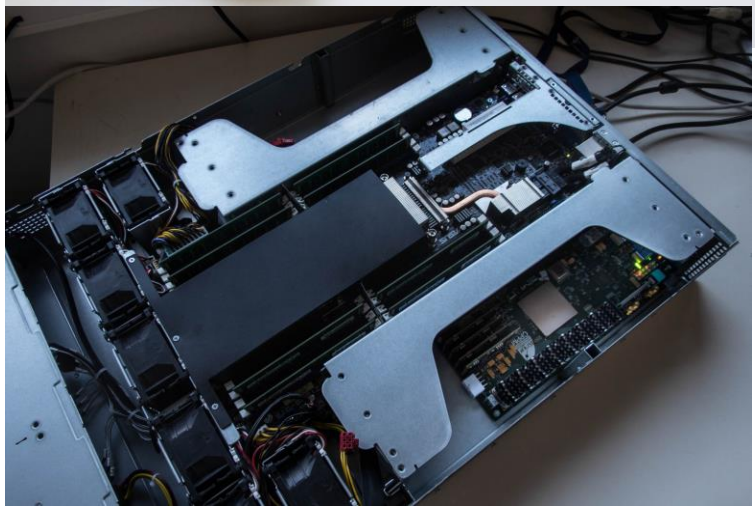
Readout boards / Event builders



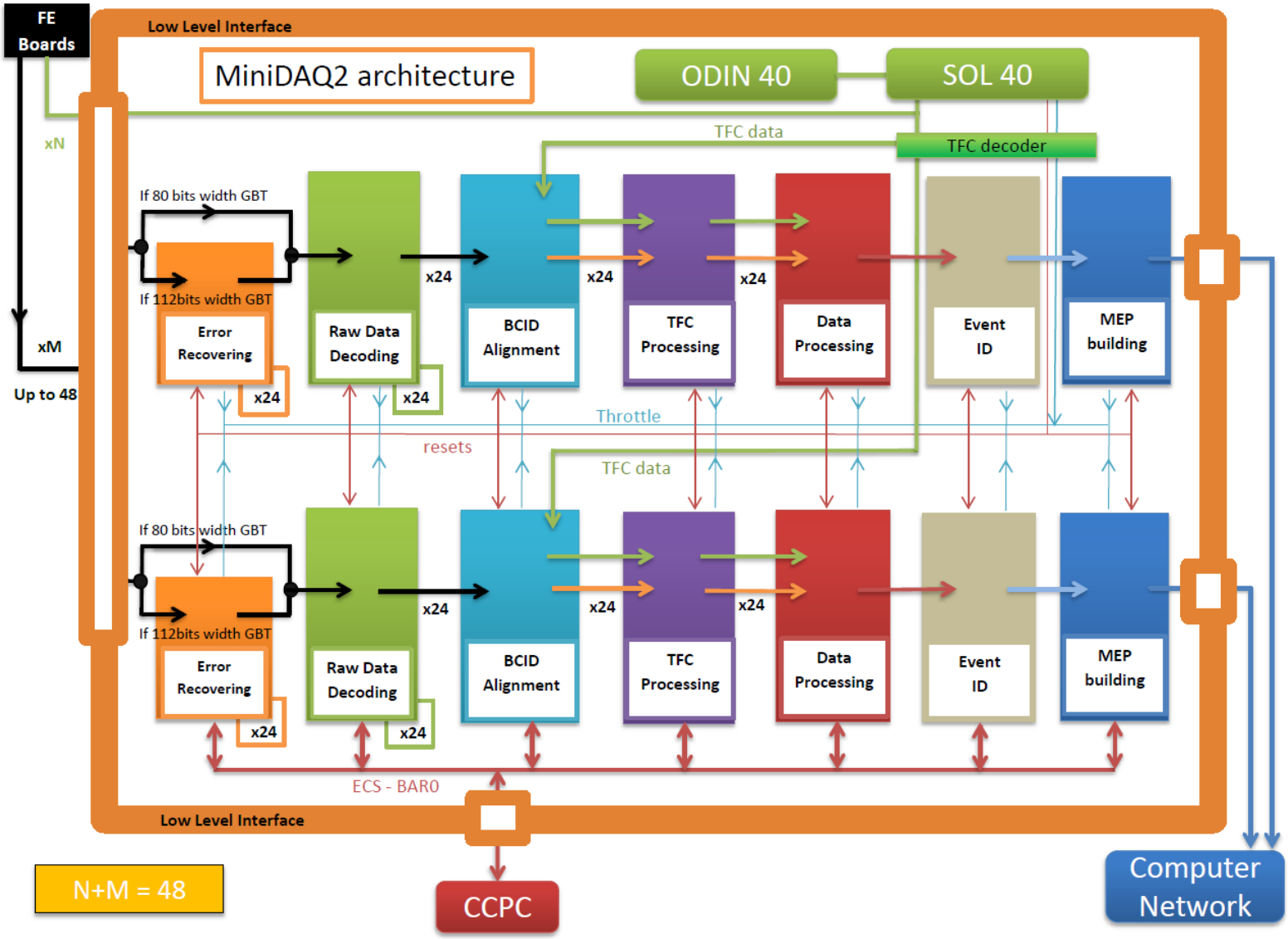
Readout board hardware (PCIe40)

- PCI Express add-in card
  - Altera Arria10 FPGA
  - 100 Gbps DMA engine to event-builder memory
- High-density optical IO
  - Up to 48 transceivers (Avago MiniPODs)
  - Reuse same HW for timing distribution system
- Decouple FPGA from network
  - Maximum flexibility in network technology
- Exploit commercial technologies
  - PCI Express Gen3 interconnect
  - COTS servers designed for GPU acceleration
- 2nd generation readout board
  - Developed at CPP Marseille
- Pre-production launched
  - Ready end of Q2
- Market survey completed
  - Tender in H2 2016
- More in the talk by P. Durante



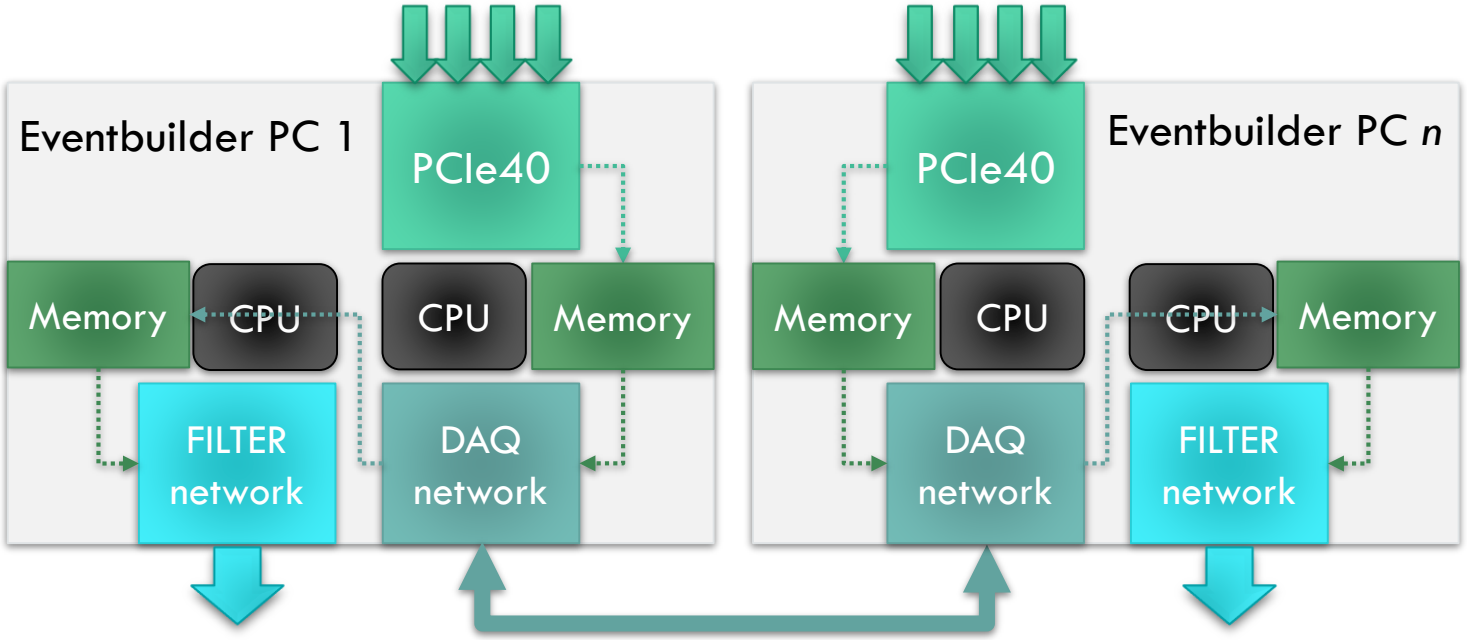


Common PCIe40 firmware frame-work

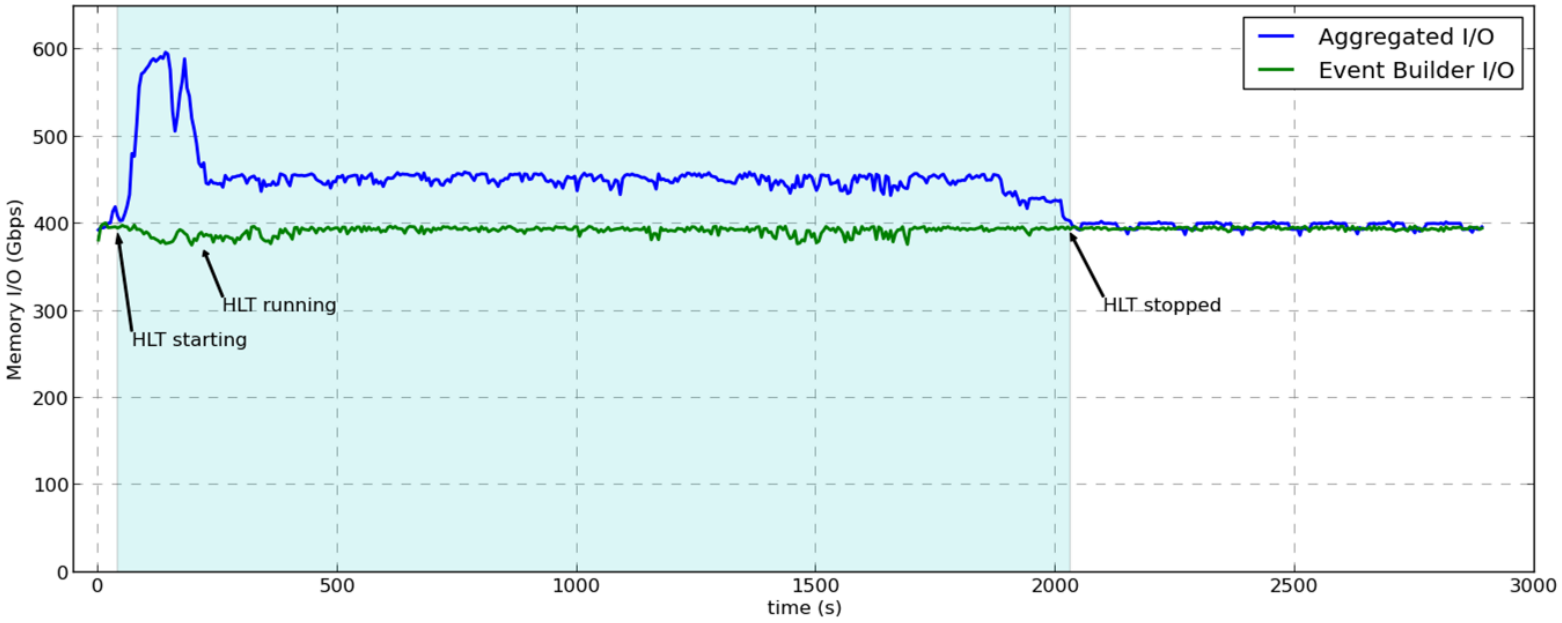


Readout unit dataflow

- A single Readout unit must sustain ~400 Gbps IO bandwidth
- Optimize memory bandwidth
  - Design for zero-copy operations and RDMA over the network
  - Organize dataflow according to topology and IO resources
  - Exploit full network bandwidth



Folded event-builder network



Cons: Every node is readout and builder unit at the same time → 400 Gbit/s I/O

Limiting factor will be the memory bandwidth (not a problem)

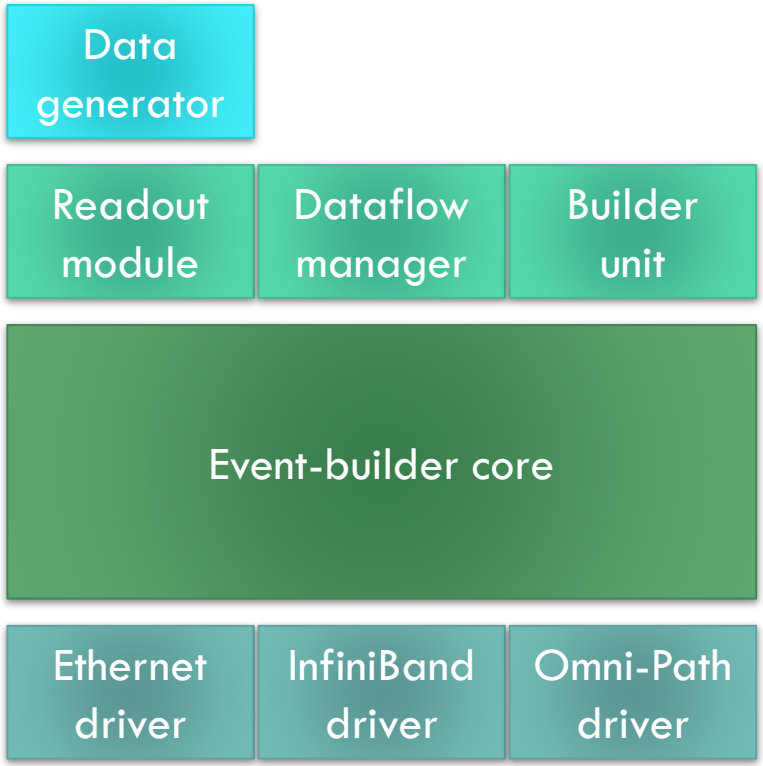
Pros: Fewer switch ports

Eventbuilder nodes see full events → opportunistic usage of the CPU

Protocol conversion: eventbuilder and filter LAN can be different technology

Event-building software (DAQPIPE)

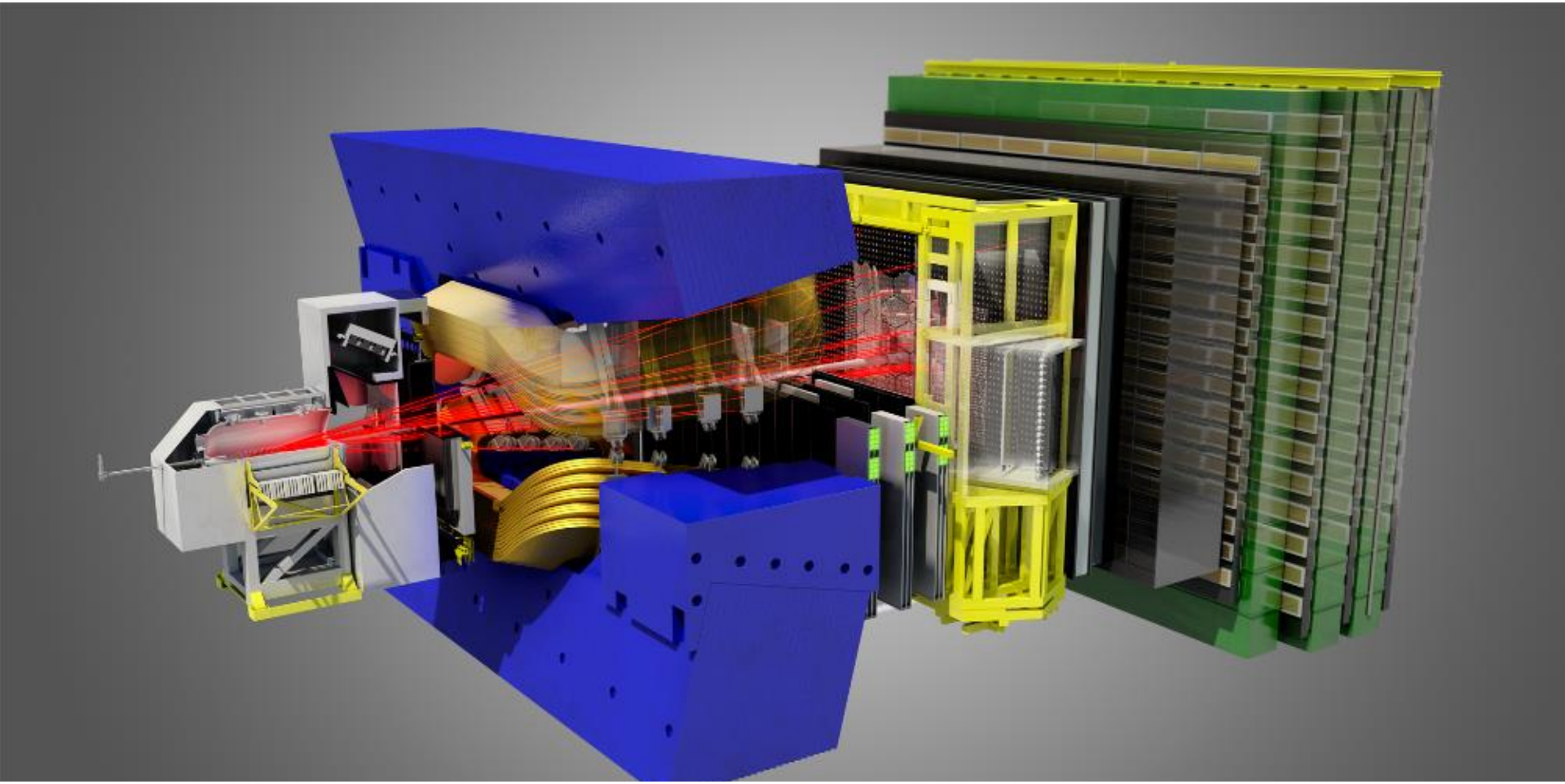
- Recreate distributed event-building dataflow of LHCb Run3
- Modular architecture, "drivers" for each network technology under evaluation
- Leverage existing HPC sites to assess scalability
- Close collaboration with the industry through CERN OpenLab
- Already achieving ~86 Gbps with Infiniband EDR and Intel OPA
  - Meets our target
  - Reduced scale setup
- External large-scale tests for InfiniBand EDR and Intel Omni-Path being prepared at HPC sites



Future data centre at Point 8

	Turnkey commercial solution (Requires minimal support from CERN engineering groups)	**Leverage existing infrastructure at Point 8**
Building	Buy pre-fabricated containers from a commercial supplier	Accommodate the farm in an existing building (SX8 hall)
Cooling	Cooling solution depends on the vendor (e. g. free air cooling)	Passive rear-door heat exchangers using primary water from existing cooling towers ▪ Compatible with DCLC for hot spots ▪ Test setup at the pit to evaluate performance with on site "warm" cooling water

A review to decide the most cost-effective solution will take place in April.



This will be the highest throughput data acquisition system ever built

Discussion slides

Is it risking to become obsolete?

The DAQ depends only on:

1. Some kind of "server" capable of housing an FPGA based electronics board
2. A local area network technology

The DAQ is scalable horizontally and vertically

PCIe roadmap is compatibly safe until at least Run4 (Gen4 available probably from late 2017 onwards)

No dependence on exact server architecture, CPU architecture, GPGPUs, specific LAN technology

Can we "unfold" the event-builder

In this case event-receiver and event-builder run on different PCs → go back to Run2 DAQ

Loose CPU power in event-builder PCs

But distribute the I/O better

Need more fully connected network ports

Again, cost will decide